Predicting soil properties in the Canadian boreal forest with limited data: comparison of spatial and non-spatial statistical approaches

Julien Beguin[1], Geir-Arne Fuglstad[2], Nicolas Mansuy[1] David Paré[1]

[1] Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, Quebec, QC G1V 4C7, Canada.

[2] Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway.

Corresponding author: J. Beguin (julien.beguin@canada.ca); Tel: +14186487414

Number of tables: 2
Number of figures: 6
Number of appendix: 2
Number of words: 8110 (including everything excepted captions of tables and figures)

ABSTRACT

Digital soil mapping (DSM) involves the use of georeferenced information and statistical models to map predictions and uncertainties related to soil properties. Many remote regions of the globe, such as boreal forest ecosystems, are characterized by low sampling efforts and limited availability of field soil data. Although DSM is an expanding topic in soil science, little guidance currently exists to select the appropriate combination of statistical methods and model formulation in the context of limited data availability. Using the Canadian managed forest as a case study, the main objective of this study was to investigate to which extent the choice of statistical method and model specification could improve the spatial prediction of soil properties with limited data. More specifically, we compared the cross-product performance of eight statistical approaches (linear, additive and geostatistical models, and four machine-learning techniques) and three model formulations ("*covariates only*": a suite of environmental covariates only; "*spatial only*": a function of geographic coordinates only; and "*covariates + spatial*": a combination of both covariates and spatial functions) to predict five key forest soil properties in the organic layer (thickness and C:N ratio) and in the top 15 cm of the mineral horizon (carbon concentration, percentage of sand, and bulk density). Our results show that 1) although strong differences in predictive performance occurred across all statistical approaches and model formulations, spatially explicit models consistently had higher R^2 and lower RMSE values than non-spatial models for all soil properties, except for the C:N ratio; 2) Bayesian geostatistical models were among the best methods, followed by ordinary kriging and machine-learning methods; and 3) comparative analyses made it possible to identify the more performant models and statistical methods to predict specific soil properties. We make modeling tools and code

available (e.g., Bayesian geostastical models) that increase DSM capabilities and support existing efforts towards the production of improved digital soil products with limited data.

Keywords: digital soil mapping, boreal forest, spatial autocorrelation, Bayesian analyses, machine-learning, random forests, boosted regression trees, kriging, geostatistic, cross-validation.

HIGHLIGHTS:

- Little guidance exists for selecting statistical models and methods in DSM with limited data.
- We performed quantitative comparisons among a range of statistical models and methods.
- Spatially explicit statistical models usually performed better than non-spatial models.
- Bayesian geostatistical models performed the best, followed by ordinary kriging and machine-learning methods.
- Our study provides modeling tools and guidance that further improve DSM capabilities.

1. Introduction

Spatially explicit soil information is required to assess potential land use, predict vulnerabilities and implement biogeochemical models forecasting the impact of human activity and climate change on terrestrial ecosystems, as well as on the services they provide (Adhikari and Hartemink, 2016; Folberth et al., 2016). Considerable efforts have been made by the research community to harmonize and define common specifications of soil data sets from different origins (Arrouays et al., 2014). These efforts have led to the creation of large soil pedon databases that facilitate the mapping, monitoring and modeling of ecosystem processes at multiple spatial scales, making it possible to predict vegetation shifts (Kuhn et al., 2015) and changes in ecosystem productivity (Maire et al., 2015). Key outcomes of these advances culminated in the release of soil raster products at continental (Hengl et al., 2015) and global (Hengl et al., 2014) scales, together with quantitative estimates of uncertainty associated with predicted soil properties. The availability of soil quantitative estimates is a significant step toward integrating soil indicators into the assessment of ecosystem function and vulnerability (Folberth et al., 2016).

Digital soil mapping (DSM) involves the use of numerical methods to fit and validate statistical models on georeferenced soil information (dependent variables) using environmental covariates (independent variables) that represent soil-forming factors, and to map predictions and their uncertainty at a specified spatial resolution over a focal study area. Environmental covariates are obtained from various sources, including remote sensing products and digital elevation models (McBratney et al., 2003). When detailed expert-based soil maps are available, techniques of spatial disaggregation of polygon information are often used (Bui and Moran, 2001; Lamboni et al., 2016). However, over large regions, and more typically in forested regions,

134 expert-based soil maps are often unavailable and the *scorpan* model approach (see McBratney et
135 al., 2003), which matches environmental covariates with soil point data (pedons), is commonly
136 used. A key challenge in DSM is that there is almost always a shortage of soil pedon data, which
137 may lead to low model accuracy and/or misrepresentations of predicted soil attributes (Ahrens et
138 al., 2008). How to make the maximum use of sparse data is thus a recurrent challenge in soil
139 science. At the same time, this challenge offers a growing opportunity to develop new statistical
140 approaches that improve soil predictive mapping.

141 Canada's forests, which cover over 390 million ha of land and represent 10% of the
142 world's forest cover, are representative of this situation since only limited soil pedon database are
143 available at the national level. Over the last decades, the pool of available numerical methods
144 and statistical models combined with an increase in computing power and data availability have
145 tremendously boosted DSM capabilities with limited data (McBratney et al., 2003; Grunwald,
146 2009; Brevik et al., 2016). Various new modeling tools are now freely available to predict and
147 map soil types as well as continuous or discrete soil properties. The quality of these predictive
148 soil maps, however, remains highly variable and depends of the interplay among four main key
149 components: 1) the availability and quality of the data for both soil profiles and environmental
150 covariates; 2) the inherent variation in nature complexity and heterogeneity of any focal soil
151 property across spatial scales and soil depth; 3) the specification of statistical models (e.g., the
152 choice of covariates, with linear vs non-linear effects, with simple vs interaction effect terms,
153 with hierarchical structures or not, with the inclusion or not of a spatial component); and 4) the
154 choice of statistical framework (e.g., Bayesian vs frequentist), statistical method and algorithm to
155 fit these models (see Fig. 1), hereafter referred as ‘statistical methods’.

Machine-learning techniques, in particular, have become very popular in predictive modeling (Hastie et al., 2009; Kuhn and Johnson, 2013), especially in DSM where numerous studies use random forests (Grimm et al., 2008), boosted regression trees (Grinand et al., 2008), *k*-nearest neighbors (Mansuy et al., 2014), Cubist (Rizzo et al., 2016), support vector machines (Were et al., 2015), and artificial neural networks (Behrens et al., 2005). In addition the choice of of statistical method, the choice of statistical models (model specification) includes the use of non-spatial vs spatially explicit models. When the geographical locations of sample plots are recorded, spatially explicit models are often used to account for spatial autocorrelation in the data or in model residuals (McBratney et al., 2003; Dormann et al., 2007; Hengl 2009; Beale et al., 2010; Banerjee et al., 2014), which often improves the accuracy of predictions as well as the predictive performance of the models (Beguin et al., 2012).

Although every spatial statistical method has its intrinsic way of modeling spatial correlation structure in the data (Li and Heap, 2014), the following are the most common in practice: 1) using additional covariates that are parametric or non-parametric functions of the sample geographic coordinates, such as in trend surface analyses with linear or additive models (Dormann et al., 2007), in spatial filtering regression (e.g., Moran Eigenvectors; Dray et al., 2006) or in autocovariate regression (Dormann et al., 2007); 2) using spatial covariance structure in the variance-covariance matrix with parametric function (e.g., variograms), such as in generalized least squares (GLS) models (Dormann et al., 2007) or in regression-kriging (Hengl et al., 2004); 3) using weighted matrices of interactions among neighboring sites, such as in conditional (CAR) and simultaneous (SAR) autoregressive models (Banerjee et al., 2014); and 4) using Bayesian hierarchical models where effects of the covariates, spatial effects and nugget effects are combined in an additive model (Banerjee et al., 2014). Bayesian methods may be

179 computationally heavy, but there has been much recent development that makes them readily
180 usable for data sets of realistic sizes of an order of 10000 points and bigger (Sun, et al., 2012,
181 Lindgren and Rue, 2015).

182 While great efforts by the soil mapping community have led to standardized technical
183 specifications regarding the spatial entity, the assessment of soil properties to be predicted, and
184 the handling of uncertainties in DSM (Arrouays et al., 2014), less work has been done to
185 compare the relative performance of a range of statistical methods and model specifications (e.g.,
186 spatial vs non-spatial) across multiple soil properties. Most DSM studies (but see Heung et al.,
187 2016) use one or a few statistical approach(es) (Poggio et al., 2013; Nawar et al., 2015), typically
188 with one type of model specification to analyze specific soil data sets, which often makes unclear
189 the extent to which the combination of particular statistical approach and model formulation
190 influences the outcome.

191 The main objective of this study was therefore to investigate to what extent the choice of
192 statistical methods and model specification could improve the spatial prediction of forest soil
193 properties with sparse soil data. More specifically, we compared the cross-product performance
194 of eight statistical methods (linear, additive and geostatistical models, and four machine-learning
195 techniques) and three different model specifications ("*covariates only*": model fitted with a suite
196 of environmental covariates only; "*spatial only*": model fitted with only a spatial component
197 derived from geographic coordinates of the plots; and "*covariates + spatial*": model fitted with
198 both covariates and a spatial component) to predict five key forest soil properties (thickness and
199 C:N ratio in the organic layer as well as carbon concentration, percentage of sand, and bulk
200 density in the top 15 cm of the mineral horizon).
201

2. Material and methods

2.1. Study area

The study area covers 290 million ha of managed forests across Canada and extends from 52° to 138° West and from 42° to 60° North (Fig. 2). The study area encompasses broad geographical gradients of topography, climate and vegetation (see details in Table 2) that influence soil formation. To predict soil properties and compare results from different models, we used a standardized data set of georeferenced soil pits (pedon) composed of ~500 georeferenced ground plots from Canada's National Forest Inventory (NFI; Gillis et al., 2005; Fig. 2). This field soil data set is similar to the one used in Mansuy et al. (2014). Our analysis focused on upland forests and therefore did not cover areas dominated by wetlands and non-forested areas (e.g., agricultural lands).

2.2. Soil data and environmental covariates

NFI soil data have the advantage of having been sampled using a standard methodology across Canada, making it representative of broad ecological conditions across the country. However, with about 500 points spaced out at least 20 km apart, NFI data have the disadvantage of being very sparsely distributed at the national scale. Soil properties in each sample were measured, recorded by depth classes, and analyzed in the laboratory according to standard protocols (Gillis et al., 2005). Among the soil attributes available in the database, we selected five soil properties, both in the organic layer and in the top 15 cm of the mineral horizon, that are commonly used as ecological indicators in natural resources management projects (Table 1). Mean, range, and coefficient of variation associated with each of the five selected soil properties are described in detail in Mansuy et al. (2014). For the statistical analysis, the soil data base was

randomly split at the beginning of the modeling process to create a training data set and an independent validation data set (Figs. 2 and 3). We tested different training-validation hold-out size combinations (90%-10%; 80%-20%; 70%-30%), but as we did not observe any difference among these combinations, we retained a 90%-10% hold-out partitioning.

We related field soil properties to 12 environmental covariates usually associated with soil formation as per the *scorpan* approach (Table 3). We first generated environmental covariates from a digital elevation model and from spatial climate models at a 250 m resolution. We retained four different topographic variables derived from the USGS/NASA SRTM 250 grid map (https://www.usgs.gov) (Table 3). We also used six climate variables derived from the spatial climate models of McKenney et al. (2011). For vegetation composition, we used derived-MODIS maps of the proportion of coniferous and deciduous species (Beaudoin et al., 2014). We also tested the use of the surficial geology map of Canada (Geological Survey of Canada, geogratis.gc.ca) as predictor, but because of a lack of explaining power, these attributes were not retained in the final analyses. We did not use soil classification maps as predictors because of their high uncertainty in Canadian boreal forests and because of the low sample size per soil class. All environmental predictors were raster layers projected into Canada's Lambert Conformal Conic with a spatial resolution of 250 m (1 pixel = 6.25 ha), and covered the entire land base of Canada's managed boreal forest (Table 3).

2.3. Statistical methods

We compared the predictive performance of eight different statistical approaches, including two parametric methods (linear and additive models), four non-parametric methods of machine learning (boosted regression trees, random forests, Cubist and weighted *k*-nearest

neighbors), ordinary and regression kriging, and one method of Bayesian inference with geostatistical models fitted with the method of integrated nested Laplace approximation (INLA) (Fig. 3). Justification for the choice of these methods is as follows: 1) except for Bayesian geostatistical models, these methods encompass the majority of statistical routines currently used in DSM studies; 2) they include algorithms that are known to be efficient in predictive modeling; 3) they encompass a gradient of complexity; and 4) they can handle spatial dependence structures in various ways (Li and Heap, 2014). Although we acknowledge that our selection of methods is not exhaustive, these methods are representative of the ones used in DSM studies and they suit the purpose of this study. For parametric approaches, we used generalized linear (GLM) and additive (GAM) models. For machine-learning approaches, we selected four supervised methods that are used extensively in predictive modeling (Kuhn and Johnson, 2013). Supervised learning encompasses a wide range of optimized algorithms designed to uncover patterns in training data, with the primary objective of making accurate predictions based on independent data (Hastie et al., 2009).

Among the four selected machine-learning methods, we retained two ensemble modeling techniques: random forests (RF) (Breiman, 2001) and boosted regression trees (BRT) (Friedman, 2001, 2002). Both techniques accommodate a variety of response types and efficiently deal with missing data and outliers, interactions among predictors, and non-linear relationships between predictors and the response variable. RF uses a modified bootstrap aggregation (bagging) algorithm (e.g., Breiman's algorithm) that fits a high number of independent classification trees on random subsamples of the original data set (Breiman, 2001). Model predictions are then estimated as weighted averages across all trees. Weights are calculated according to the predictive performance of each tree, which ensures that the best trees contribute the most to final

predictions (Hastie et al., 2009). RF differs from other decision tree methods in that each regression tree is fitted with a different random subset of covariates based on random subspace techniques (Ho, 1998). We performed extensive sensitivity analyses on two key parameters: the maximum number of trees and the number of predictors for each tree. More information on the theoretical and practical aspects of RF can be found in Breiman (2001), Liaw and Wiener (2002), and Hastie et al. (2009).

BRT combines large numbers of regression tree models adaptively to optimize predictive performances (Friedman, 2002) and differs from other decision trees techniques by using a boosting algorithm (e.g., AdaBoost: Freund and Schapire, 1996). The purpose of a boosting algorithm is to minimize a loss function averaged over random training data sets and to seek by the use of successive iterations, to improve predictions on poorly fitted data points (see Elith et al. (2008) for more details). BRT requires more tuning than RF and we performed extensive sensitivity analyses on four key parameters: tree complexity or interaction depth (*tc*), learning rate or shrinkage parameter (*lr*), number of trees (*nt*), and minimum number of observation in nodes (*minobs*) (see also Appendix 1.1). A full description of these parameters is beyond the scope of this study, but detailed information can be found in Elith et al. (2008) and Kuhn and Johnson (2013).

The third machine-learning method used in this study is Cubist, a rule-based regression technique developed by Quinlan (1992, 1993), which fits a separate multiple linear model at each leaf node of a regression tree according to a set of conditional rules (Walton 2008). These rules have the form of conditional logical statements (*if… then…*) and divide the training data set into multiple subsets, where each subset is composed of data points sharing common statistical properties. A regression model is then fitted separately for each rule/subset. In contrast with

294 other tree methods, Cubist allows for more than one conditional statement to be combined at
295 each intermediate leaf node of a tree. Predictions are then made based on the match with a rule's
296 condition and its associated regression model. To improve prediction accuracy, Cubist can be
297 coupled with a bootstrap aggregation algorithm (committees > 1), where the number of
298 committee models defines the number of boosting iterations. In addition, Cubist can incorporate
299 composite instance-based models by varying the number of nearest neighbors. More details on
300 Cubist and these options can be found here: www.rulequest.com. We performed sensitivity
301 analyses on two key parameters: the number of committees and the number of nearest neighbors
302 (see also Appendix 1.1).

303 The last machine-learning method we tested is derived from the *k*-nearest neighbors (*k*-
304 NN) method (Cover and Hart, 1967). The *k*-NN method is based on the assumption that the more
305 two observations are similar regarding their range of values for a set of independent variables,
306 the more their predicted values for a response variable of interest should be close. Predicting
307 values for a new observation can therefore be made by using the sampled observations from a
308 training data set that are the closest (= nearest neighbor(s)) to each new observation with respect
309 to the covariates used (not to be confused with nearest neighbors in geographic space).
310 Determination of the similarity between new and training samples is based on distance metrics,
311 and the method has been expanded to a set (*k*) of nearest neighbors. The weighted *k*-NN
312 (KKNN) method extends *k*-NN by adding the characteristic that sample points within the
313 training data set that are particularly similar to the new observation should have more weight in
314 the decision than neighbors that are further away from it (Hechenbichler and Schliep, 2004). The
315 KKNN method used in this study fits kernel functions to weigh nearest neighbors according to
316 their similarity distance from each new data point, and uses these weights in combination with

317 covariates to make predictions. For each predicted soil variable, we performed sensitivity
318 analyses on both the type of kernel function for weighing neighbors and the optimal maximum
319 number (= k) of nearest neighbors (see also Appendix 1.1).
320 In a Bayesian linear geostatistical model, the dependent variable is described as an
321 additive combination of latent explanatory components, such as linear effects of independent
322 variables, a spatial effect indexed by the geographical locations, and a nugget effect representing
323 a discontinuity from the semi-variogram at its origin. The main advantages of such a model over
324 the machine-learning methods are 1) the interpretability gained by an explicit model as to how
325 observations are generated; and 2) in the Bayesian framework, uncertainty estimates are directly
326 available for all parameters as well as for predictions. The disadvantages of Bayesian methods
327 are that they can be difficult to implement and are computationally expensive to run. The INLA
328 methodology (Rue et al., 2009) combined with the recent stochastic partial differential equation
329 (SPDE) approach to spatial fields (Lindgren et al., 2011), provides a methodological solution to
330 these limiting factors. Moreover, the INLA R-package makes them easy to use (Lindgren and
331 Rue, 2015). The main difference between the INLA methodology and the standard use of
332 Markov Chain Monte Carlo (MCMC) simulations is that INLA is not based on sampling, but
333 rather on deterministic approximations to integrals, which makes it computationally quicker than
334 MCMC approaches, while still being very accurate (Rue et al., 2009).
335
336 *2.4. Model specifications*
337 To evaluate if patterns of spatial autocorrelation among sample plots could be used as
338 surrogates for unmeasured covariates and contribute to explaining more residual variation in soil
339 properties than the variation explained by environmental covariates alone, we fitted and

compared, for each statistical method, three different statistical model specifications (Fig. 3): 1) a first type of model (labelled "*covariates only*") that uses only environmental covariates (topography, vegetation, and climatic conditions) as predictors. This type of model is non-spatial because no explicit spatial relationships among sample plots are quantified; 2) a second type of model (labelled "*spatial only*") that only uses a spatial component, i.e. a function of the geographic coordinates of sample plots (latitude and longitude), to predict soil properties. Note that the specification of spatial functions differs for each statistical method (see section 2.5.). However, all "*spatial only*" models share the characteristic that they only consider spatial locations and distances among plots to predict and map soil properties; 3) the last type of model (labelled "*covariates + spatial*") is a combination of the two first types, in which soil properties are predicted as a function of both environmental covariates and a spatial component that quantifies possible spatial relationships among sample plots.

2.5. Spatially explicit models

As the way of quantifying spatial relationships among sample plots varies according to the statistical method used, we used method-specific spatial functions in the "*spatial only*" and "*covariates + spatial*" models that belong to one of these three classes: kriging, stochastic partial differential equations, or models that use a function of spatial coordinates (latitude + longitude) of sample plots directly as predictors. For kriging, we compared local ordinary kriging, and regression kriging both fitted with an exponential variogram model. We assumed stationarity and isotropy in all cases. For local ordinary kriging ("*spatial only*" model), we set the maximum number of nearest neighbors to 20. In regression kriging ("*covariates + spatial*" model), a regression model was first fitted to the data using environmental covariates as predictors, and

ordinary kriging was performed in a second step on the model's residuals (Hengl et al., 2004). Kriged residuals were then added to the predictions of the regression model. In this study, we compared two regression kriging methods: one in which the relationship between each soil property and environmental covariates is evaluated through a GLM (Hengl et al., 2014), and the other in which the relationship between each soil property and environmental covariates is evaluated through the RF algorithm (Hengl et al., 2015).

For spatial GAMs, we modelled spatial relationships as covariates using a trend-surface obtained from a two-dimensional spline function on geographical coordinates (Dormann et al., 2007). In addition to the RF-kriging described above, we fitted another spatial RF model using the geographical coordinates of soil samples directly as predictors along with the other environmental covariates. As the RF algorithm automatically evaluates interactions among covariates, the use of geographical coordinates as covariates allowed us to account for possible latitudinal and/or longitudinal gradients in the effect of environmental covariates on soil properties. We used the same approach, with the inclusion of geographical coordinates of the samples as predictors, for BRT, Cubist, and KKNN. For completeness and comparative purposes, we fitted a linear model with only the latitude and longitude of the sample plots as simple and linear effect.

For Bayesian geostastistical models, we used INLA together with the SPDE approach proposed by Lindgren et al. (2011), where the spatial field is described as the solution to a linear differential equation driven by white noise and involves the construction of meshes on which the spatial field is defined (see Blangiardo and Cameletti, 2015; Appendix 1.3). With the Bayesian framework, priors need to be defined for each parameter of the model. We used non-informative default priors for coefficients of the environmental covariates, and a weakly informative prior for

386 the spatial effect (Fuglstad et al., 2015; see Appendix 1.3 for the definition of priors associated
387 with the spatial effect and the variance of the nugget effect). Overall, the mixture of eight
388 statistical methods and three different model specifications yielded a total of 24 unique method-
389 model combinations for each soil property (see Fig. 4).
390
391 *2.6. Evaluation of predictive performances*
392 All models were calibrated on 90% of the soil data base using cross-validation and
393 validated on the remaining 10% of the soil data base (see Fig. 3). To address the fundamental
394 trade-off between model complexity and prediction errors on independent data sets (see p. 220 in
395 Hastie et al., 2009), we pruned all our models during the calibration stage and only retained
396 relevant environmental covariates that minimized root mean square error (RMSE) values using
397 repeated 10-fold cross-validation. We made sure that comparisons among the three types of
398 model ("*covariates only*", "*spatial only*", and "*covariates + only*") for each statistical method
399 were valid by retaining the same set of environmental covariates for every predicted soil property
400 (see Table 1). For all models, we calculated the coefficient of determination (R^2) and RMSE
401 using 10-fold cross-validation repeated 20 times (see Fig. 3) (Bennett et al., 2013). This standard
402 procedure prevented our analyses from having overfitting issues, while providing fair estimates
403 of prediction errors (Hastie et al., 2009). Cross-validated R^2 was calculated as $R^2 = 1 - ESS/TSS$,
404 where:
405 $$ESS = \sum_i (y_{i\,(observed)} - y_{i\,(predicted)})^2$$
406 $$TSS = \sum_i (y_{i\,(observed)} - \bar{y}_{(observed)})^2$$
407 We report the quantile distribution (0.025%, 50%, and 97.5%) of cross-validated R^2 and
408 RMSE calculated from 20 runs, which gives us information on the variation in mean cross-

validated R^2 and RMSE across repetitions. We reported cross-validated RMSE values as percentages of the sample mean (RMSE (%) = 100 x RMSE/$\bar{y}$). Finally, to ensure independent validation, we validated all our models in calculating independent R^2 between predictions from cross-validated models with observed values from independent data (Figs. 3 and 5). All models were run in the R environment (R Core Team, 2015). INLA was run using the '*R-INLA'* package (Rue et al., 2009), ordinary kriging was fitted using the '*gstat'* package (Pebesma and Graeler, 2013), and we used the '*GSIF'* package for regression-kriging and RF-kriging (Hengl et al., 2016). All other statistical methods and models were run using the *'caret'* meta-package (Kuhn, 2015) in combination with the following R packages: *'gbm'* (Ridgeway 2015: BRT), *'RandomForests'* (Liaw 2015), *'Cubist'* (Kuhn et al., 2015), *'kknn'* (Schliep and Hechenbichler, 2015), and *'mgcv'* (Wood, 2015; GAM). All R-codes used to run our analyses are available in Appendix 1.

3. Results

3.1. Performances of statistical methods and models

Our comparative analyses revealed that the choice of the statistical method, the type of model specification ("*covariates only*", "*spatial only*", or "*covariates + spatial*"), or their interaction significantly influenced the predictive performance for most soil properties tested in this study (Figs. 4 and 5). For instance, cross-validated R^2 and RMSE values of non-spatial models for sand content (%) were respectively over 100% higher and lower when using either RF or BRT (median cross-validated R^2 ranging from 0.23 to 0.28; Fig. 4) than when using either linear or additive models (median cross-validated R^2 ranging from 0.04 to 0.13; Fig. 4). On the other hand, the best cross-validated R^2 and RMSE values of "*spatial only*" models for sand

content were obtained with INLA/SPDE, followed by ordinary kriging and machine-learning techniques (median cross-validated R^2 ranging from 0.38 to 0.46; Fig. 4). For "*covariates + spatial*" models, INLA/SPDE yielded the best predictive performance for sand, followed by machine-learning techniques and linear regression-kriging (Fig. 4).

Although strong differences in predictive performance occurred across all statistical methods and model types, models that incorporated a spatial component consistently had higher cross-validated R^2 and lower cross-validated RMSE values than non-spatial models for all soil properties, except for the C:N ratio in the organic layer (Fig. 4). This result indicates that the spatial relationships among soil samples contain valuable information that is not captured by environmental covariates and that the inclusion of spatial information improves the overall predictive performance (Fig. 4). This improvement was also verified on independent data sets (Fig. 5). Interestingly, "*covariates + spatial*" models outperformed "*spatial only*" models for only one soil property, the organic layer C:N ratio (Fig. 4). This result highlights that in most cases in our study, there is a significant redundancy between the variation captured either by environmental covariates or by spatial functions.

3.2. Mapping predictions

Mapping the predicted values (mean + standard deviation) from "*spatial only*" or "*covariates + spatial*" models using machine-learning methods with latitude and longitude of soil samples as covariates occasionally generated spatial discontinuities with sudden transition patterns along longitude in predicted surfaces (results not shown). Such prediction artefacts, which could not be detected by analyzing the fit of the model alone, did not occur with spatial models fitted with INLA or with regression kriging. When spatial relationships among soil

samples explained a significant part of the total variance (Fig. 4), the spatial distribution of prediction errors also exhibited strong spatial patterns (Fig. 6D for sand), with clusters containing the highest level of uncertainty in areas with low sampling density. Inversely, when the results showed low levels of residual spatial structure, such as for the C:N ratio in the organic layer (Fig. 4), the spatial distribution of prediction errors was more homogeneous across landscapes (see Fig. 6E for the C:N ratio). These results may be used to improve data acquisition strategies.

4. Discussion

Our main findings demonstrate that, irrespective of the effects of environmental covariates and of the inherent variation present in soil profile data sets, both the choice of statistical method and the choice of model type can have a significant impact on the predictive performance for most predicted forest soil properties tested in this study. These results are inconsistent with the view that the choice of statistical method and the type of model specification would have a negligible influence on the performance and accuracy of predicted soil maps calibrated with limited soil data. Our study shows that selecting one suboptimal combination of statistical method and model type could lead to a decrease in predictive power as high as 100% and sometimes higher (see Fig. 4). These results have a number of important implications for further work on DSM in a context of limited soil data.

Our results indicate that most forest soil properties are characterized by some degree of spatial autocorrelation that may have different degrees of redundancy with the information contained in the environmental covariates available. If the spatial information contained in soil data points overlaps strongly with the one contained in the environmental covariates, including

both sorts of information in a same model may lead to unnecessary complexity in the model. As predicted by the bias-variance trade-off, models that are overfitted on training data often have higher prediction errors with independent data (Hastie et al. 2009). Our results on sand, thickness and bulk density are in agreement with these predictions as more complex "*covariates + spatial*" models fitted with machine learning methods often had lower cross-validated and independent R^2 (Figs. 4 and 5) than "*spatial only*" models. Interestingly, such a pattern did not occur with the use of the SPDE/INLA approach where either "*covariates + spatial*" or "*spatial only*" models were very similar in terms of predictive performance (see Figs. 4 and 5). At the opposite, if the information captured by spatial functions and environmental covariates is additive and contains low levels of redundancy, a model formulation including both the effects of environmental covariates and an appropriate spatial structure will be useful to decrease prediction errors on independent data. This raises the question as to whether or not spatial models should be systematically tested in DSM exercises when soil pit data are limited? A pragmatic answer would be 'yes', but more studies are needed in this area as the benefits of including spatial correlation structures may also depend on the sampling efforts or the breadth of environmental gradients present in the study area and in the sampled data.

Under our study conditions, more specifically a very large land base, a limited number of soil pit data and a high minimal distance between soil pits (> 20 km), "c*ovariates only*" models outperformed "*spatial only*" models for only one of the soil variables studied: the C:N ratio of the organic layer. Also, only for this soil property did the "*covariates + spatial*" models clearly outperform "*spatial only*" models. A possible explanation for this result is that forest composition, which is known to influence soil C to N stoichiometry (Averill et al. 2014), mostly varies over distances < 20 km because of its strong dependence on local disturbance history

(Chen et al., 2009). A distance variation < 20 km could not be captured accurately by the sampling scheme of our data set; hence, the consideration of covariates expressing forest composition appears to bring relevant information that is non-redundant with that captured by the "*spatial only*" models. As for the other soil properties to be predicted, the limited improvement obtained by adding environmental covariates, once spatial information has been taken into consideration, indicates that the environmental covariates considered either contributed little to explaining these soil properties, or that their explaining potential is redundant with the spatial information contained in the data. Such was the case for climate variables that varied over large distances (> 100 km) and contributed to predict several forest soil properties in "*covariates only*" models, but were redundant in "*spatial only*" and "*covariates + spatial*" models once spatial correlation structures had been taken into account. Although it can be challenging to identify which ecological process(es) is(are) the cause(s) of the observed patterns of spatial autocorrelation, the use of spatially-explicit statistical models, together with cross-validation techniques, may at least provide practical alternatives to improve the accuracy of and decrease prediction errors in digital soil maps calibrated with limited data.

The digital soil maps obtained with the best model fitted with INLA showed patterns that are consistent with the large-scale gradients observed in existing national soil information products (Canadian System of Soil Classification: http://sis.agr.gc.ca/cansis/). For example, the map of sand content (Fig. 6A) highlights the fine texture soils of the Ontario-Quebec Clay Belt region, formed by the draining of the former proglacial Lake Ojibway around 8,200 BP (Vincent and Hardy, 1977), as well as the large offshore glaciolacustrine sediment of northern Manitoba (http://www.gov.mb.ca/iem/geo/). The coarse texture soils, abundant in northern Saskatchewan (http://publications.gov.sk.ca/) are also apparent. The map of organic layer C:N ratio showed a

latitudinal gradient with highest values in the north and lowest ones in the south, which is closely associated with the percentage of coniferous species in the canopy of Canadian forests (Beaudoin et al., 2014). The spatial distribution of bulk density is more complex as it relates to several soil properties, including soil texture, organic matter content, and soil compaction (Sequeira et al., 2014). The use of the most performant models substantially improved the goodness-of-fit obtained from previous national modeling efforts for sand content (+ 131 %), bulk density (+ 105 %), thickness of the organic layer (+ 205 %), and C:N ratio in the organic layer (+ 596 %) (Mansuy et al., 2014).

Another implication of our results is that the performance and ranking of modeling strategies may also depend on the property to be predicted, highlighting the need to move beyond the single statistical model–method philosophy for DSM. Clearly, more testing and quantitative comparisons are needed to get a comprehensive picture of the model-method combination that is best in specific situations. A significant step forward would be the completion of a comparative modeling study of soil properties in which the efficiency of a wide range of combinations of statistical methods and model classes would be evaluated over multiple spatial scales, along a gradient of sampling density, and across contrasting ecosystems. It would also be relevant that such a synthesis be expanded to statistical methods not commonly used in soil research. For instance, the use of Bayesian linear geostatistical models consistently yielded some of the best predictive performances across the soil properties tested in this study. Although Bayesian geostatistical models have been used with success in a wide range of applied contexts in environmental sciences (Rue et al., 2016), it is surprising to note that their use in DSM has been very limited so far. As our results point out, this might partly be due to the fact that these models come with an increase in computational time (Appendix 2), require substantial knowledge in

programming, and involve a relevant choice of priors. However, technical guides available to non-expert programmers now exist to facilitate the implementation of these types of model (Blangiardo and Cameletti, 2015), and we freely provide all the scripts needed to replicate our analyses with k-fold cross-validation and mapping procedures in the R environment. The free availability of statistical routines to perform comparative analyses in open-access, efficient and repeatable ways opens new perspectives in this area and provides strong capabilities to non-expert programmers or to researchers working in environments where programming expertise is limited.

5. Conclusion

Results from our comparative study suggest that better mapping of soil properties may be achieved through quantitative assessment when selecting the statistical method and model specification (spatial vs non-spatial models). Overall, spatially explicit models showed significantly better predictive performance (R^2 and RMSE), with improvements ranging from 10% to more than 100% compared with non-spatial models, depending on the soil property. Bayesian geostatistical models fitted with INLA showed among the best predictive performances and mapping properties with our data set. The use of comparative statistical analyses as a standard modeling practice, which goes beyond the single model–statistical method philosophy, would be a valuable asset to increase the development of DSM capabilities. Therefore, this study constitutes a step forward in the improvement of DSM capabilities by providing a quantitative assessment of the performance of a variety of spatial techniques coupled with systematic comparisons with non-spatial models. We provide scripts and suggestions for facilitating the achievement of this standard in applied research communities with limited computational

resources and programming expertise. Finally, to avoid overfitting issues and to make comparison among DSM studies possible, we would recommend that the methodology for model calibration and validation in DSM studies be better standardized and applied rigorously, for example, by using repeated k-fold cross-validation.

Acknowledgments

This study was funded through a grant to David Lee (Agriculture and Agri-Food Canada) titled *Enhancing the Biomass Inventory Mapping and Analysis Tool by Integrating and Updating Forestry, Agriculture, and Municipal Solid Waste Data and Associated Sustainability Information* of Natural Resources of Canada's Program of Energy Research and Development (PERD). The authors thank André Beaudoin and Philippe Villemaire for access to vegetation maps and Luc Guindon for discussions on mapping discontinuities. We sincerely thank Isabelle Lamarre for English editing and three anonymous reviewers for improving a previous version of the manuscript.

References

Adhikari, K., Hartemink, A.E., 2016. Linking soils to ecosystem services—A global review. Geoderma 262:101-111.

Ahrens, R.J. 2008. Foreword. In: Hartemink, A.E., McBratney, A.B., de Lourdes Mendonça-Santos, M., (Eds.) 2008. Digital soil mapping with limited data. Springer Science & Business Media, Berlin. Pp v-vi.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., Mckenzie, N.J., Mendonca-santos, M.D.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.-L., 2014. GlobalSoilMap: toward a fine-resolution global grid of soil properties. Adv. Agron. 125:93-134.

Averill, C., Turner, B.L., Finzi, A.C., 2014. Mycorrhiza-mediated competition between plants and decomposers drives soil carbon storage. Nature 505(7484):543-545.

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2014. Hierarchical modeling and analysis for spatial data. CRC Press, Boca Raton, FL.

Beaudoin, A., Bernier, P.Y., Guindon, L., Villemaire, P., Guo, X.J., Stinson, G., Bergeron, T., Magnussen, S., Hall, R.J., 2014. Mapping attributes of Canada's forests at moderate resolution through k-NN and MODIS imagery. Can. J. For. Res. 44(5):521-532.

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J., Elston, D.A., 2010. Regression analysis of spatial data. Ecol. Lett. 13:246–264.

Beguin, J., Martino, S., Rue, H., Cumming, S.G. 2012. Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation. Methods Ecol. Evol. 3(5):921-929.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. Soil Sci. 168(1):21-33.

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S. H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Software 40:1-20.

Blangiardo, M., Cameletti, M., 2015. Spatial and spatio-temporal bayesian models with R-INLA. John Wiley & Sons, New York.

Breiman, L., 2001. Random forests. Machine learning 45(1):5-32.

Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A,. 2016. Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma 264:256-274.

Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103(1):79-94.

Chen, H.Y.H., Vasiliauskas, S., Kayahara, G.J., Ilisson, T., 2009. Wildfire promotes broadleaves and species mixture in boreal forest. For. Ecol. Manag. 257:343–350.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13(1): 21-27.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kuhn, I., Ohlemuller, R., Peres-Neto, P.R., Reineking, B., Schroder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:609–628.

Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecol. Model. 196(3):483-493.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77(4):802-813.

Folberth, C., Skalský, R., Moltchanova, E., Balkovič, J., Azevedo, L.B., Obersteiner, M., van der Velde, M., 2016. Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. Nat. Comm. 7: doi:10.1038/ncomms11872

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29:1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38:367–378.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. ICML 96:148-156.

Fuglstad, G.A., Simpson, D., Lindgren, F., Rue, H., 2015. Constructing Priors that Penalize the Complexity of Gaussian Random Fields. Cornell University Press, Ithaca, NY. arXiv preprint arXiv:1503.00256.

Gillis, M.D., Omule, A.Y., Brierley, T., 2005. Monitoring Canada's forests: The national forest inventory. For. Chron. 81(2):214-221.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using Random Forests analysis. Geoderma 146(1):102-113.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143(1):180-190.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152:195–207.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction. Springer, New York.

Hechenbichler, K., Schliep, K., 2004. Weighted k-nearest-neighbor techniques and ordinal classification. Discussion paper 399, Sonderforschungsbereich 386, Ludwig-Maximilians-Universität München, Munich.

690 Hengl, T., 2009. A practical guide to geostatistical mapping, 2nd Ed. EU Publications Office,
691 Luxembourg.
692
693 Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction of soil
694 variables based on regression-kriging. Geoderma 120(1):75-93.
695
696 Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B., Ribeiro, E., Samuel-
697 Rosa., A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014.
698 SoilGrids1km—global soil information based on automated mapping. PloS ONE
699 9(8):e105992.
700
701 Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., Sila, A.,
702 MacMillan, R.A., de Jesus J.M., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250
703 m resolution: Random forests significantly improve current predictions. PloS ONE
704 10(6):e0125814.
705
706 Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview
707 and comparison of machine-learning techniques for classification purposes in digital soil
708 mapping. Geoderma 265:62-77.
709
710 Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans.
711 Pattern Anal. Machine Intel. 20(8):832-844.
712
713 Kuhn, E., Lenoir, J., Piedallu, C., Gégout, J.C., 2016. Early signs of range disjunction of sub-
714 mountainous plant species: an unexplored consequence of future and contemporary climate
715 changes. Global Change Biol. 22:2094–2105.
716
717 Kuhn, M., Johnson, K., 2013. Applied predictive modeling. Springer, New York.
718
719 Kuhn, M., 2015. The caret Package: Classification and Regression Training. R Fondation for
720 Statistical Computing, Vienna, Austria. http://caret.r-forge.r-project.org/
721
722 Kuhn, M., Weston, S., Keefer, C., Coulter, N., Quinlan, R., 2015. The Cubist Package: Rule- and
723 Instance-Based Regression Modeling. R Fondation for Statistical Computing, Vienna,
724 Austria. https://cran.r-project.org/web/packages/Cubist/index.html
725
726 Lamboni, M., Koeble, R., Leip, A., 2016. Multi-scale land-use disaggregation modelling:
727 Concept and application to EU countries. Environ. Model. Software 82:183-217.
728 Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: A
729 review. Environ. Model. Software 53:173-189.
730 Liaw, A., 2015. The RandomForest Package: Breiman and Cutler's Random Forests for
731 Classification and Regression. R Fondation for Statistical Computing, Vienna, Austria.
732 https://cran.rproject.org/web/packages/randomForest/index.html

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2(3):18-22.

Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. J. Stat. Software 63(19). doi: 10.18637/jss.v063.i19

Lindgren, F., Lindstrøm, J. Rue, H., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. Roy. Stat. Soc. Ser. B, Stat. Method. 73:423–498.

Maire, V., Wright, I.J., Prentice, I.C., Batjes, N.H., Bhaskar, R., van Bodegom, P.M., Cornwell, W.K., Ellsworth, D., Niinemets, Ü., Ordoñez, A., Reich, P.B., Santiago, L.S., 2015. Global effects of soil and climate on leaf photosynthetic traits and rates. Glob. Ecol. Biogeogr. 24(6):706-717.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method. Geoderma 235:59-73.

McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117:3-52.

McKenney, D.W., Hutchinson, M.F., Papadopol, P., Lawrence, K., Pedlar, J., Campbell, K., Milewska, E., Hopkinson, R., Price, D., Owen, T., 2011. Customized spatial climate models for North America. Bull. Am. Meteorol. Soc. 92, 1612–1622.

Nawar, S., Buddenbaum, H., Hill, J., 2015. Digital mapping of soil properties using multivariate statistical analysis and ASTER data in an arid region. Rem. Sens. 7(2): 1181-1205.

Pebesma, E., Graeler, B., 2013. The gstat Package: spatial and spatio-temporal geostatistical modelling, prediction and simulation. R package version, 1-0. R Fondation for Statistical Computing, Vienna, Austria.

Poggio, L., Gimona, A., Brewer, M.J., 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma 209:1-14.

Quinlan, J.R., 1992. Learning with continuous classes. In: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence. World Scientific Publishing, Hackensack, NY, pp. 343-348.

Quinlan, J.R., 1993 Combining instance-based and model-based learning. In: Proceedings of the 10th International Conference on Machine Learning. The international Machine Learning: 236-243.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ridgeway, G., 2015. The gbm Package: Generalized Boosted Regression Models. R Foundation for Statistical Computing, Vienna, Austria. https://cran.r-project.org/web/packages/gbm/index.html

Rizzo, R., Demattê, J.A., Lepsch, I.F., Gallo, B.C., Fongaro, C.T., 2016. Digital soil mapping at local scale using a multi-depth Vis–NIR spectral library and terrain attributes. Geoderma 274:18-27.

Rue, H, Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). J. Roy. Stat. Soc., Ser. B 71(2):319-392.

Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., Lindgren, F., 2016. Bayesian Computing with INLA: A Review. Cornell University Press, Ithaca, NY. arXiv preprint arXiv:1604.00860.

Schliep, K., Hechenbichler, K., 2015. The kknn package: Weighted k-Nearest Neighbors. R Foundation for Statistical Computing, Vienna, Austria. https://cran.r-project.org/web/packages/kknn/index.html

Sequeira, C.H., Wills, S.A., Seybold, C.A., West, L.T., 2014. Predicting soil bulk density for incomplete databases. Geoderma 213:64-73.

Sun, Y., Li, B., Genton, M.G., 2012. Geostatistics for large datasets. In: Porcu, E., Montero, J-M., Schlather, M., Advances and challenges in space-time modelling of natural events. Springer Berlin Heidelberg. pp. 55-77

Vincent, J.S., and Hardy, L., 1977. L'évolution et l'extension des lacs glaciaires Barlow et Ojibway en territoire québécois. Geogr. Phys. Quat. 31 :357–372.

Walton, J.T., 2008. Subpixel Urban Land Cover Estimation. Photogram. Engineeri. Rem. Sens. 74(10):1213-1222.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52:394-403.

Wood, S., 2015. The mgcv Package: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation". R Foundation for Statistical Computing, Vienna, Austria. https://cran.r-project.org/web/packages/mgcv/index.html.

8. Appendices

Appendix 1: Supplementary R-code for running all analyses presented in the article on the "meuse" ('sp' R-package) data set:

- Appendix 1.1: Random forests (RF), boosted regression trees (BRT), weighted-KNN (KKNN), Cubist, generalised linear (GLM) and additive (GAM) models.
- Appendix 1.2: Ordinary kriging, regression-kriging & RF-kriging models.
- Appendix 1.3: Bayesian geostatistical models with SPDE/INLA methods.

Appendix 2: Comparison of computational time among statistical methods.

TABLES

Table 1. Description of the five soil properties used in this study.

Soil horizon	Soil property	Abbreviation
Organic layer	Thickness (cm)	Thickness
	Carbon:nitrogen ratio	C:N
Mineral horizon (0-15 cm)	Carbon concentration (g/kg)	Cmin
	Sand content (%)	Sand
	Bulk density (g/cm^3)	BD

Table 2. Description of environmental covariates (rasters with pixel resolution of 250 m x 250 m) used to predict each of the five soil properties (see Abbreviations in Table 1) across the boreal forest of Canada.

				Predicted soil property				
Covariates	**Category**	**Definition**	**Mean [min;max]**	**Sand**	**C:N**	**Cmin**	**Thick -ness**	**BD**
Elevation	topography	Elevation from the Shuttle Radar Topography Mission (m)	563 [0; 3950]	√	√	√	√	√
Slope	topography	Rate of maximum change in elevation from each pixel (%)	4.2 [0; 64.8]				√	
Beers aspect	topography	Heat index = (1+cos(45-Aspect))/Slope	0.99 [0-2]	√				
Watersheld structure	topography	Local drainage area enclosed between the local divide and the stream into which each cell drains	483 [-30; 2582]					√
Acmi	climate	Annual moisture index (cm/year)	43 [-127; 385]	√				
Scmi	climate	Summer moisture index (cm/summer)	-2.5 [-76.5; 676]	√				
Pwq	climate	Precipitation of the warmest quarter (mm)	99 [0-796]	√				
Tcm	climate	Lowest temperature of any monthly minimum (°C)	-12.5 [-48.9; 6.1]			√		√
Thm	climate	Highest temperature of any monthly maximum (°C)	11.2 [-4.5; 37.8]					√
Tap	climate	Total annual precipitation (mm)	315 [0; 4302]				√	
Deciduous	vegetation	Percentage of deciduous species in the pixel (%)	19.0 [0; 100]	√	√	√	√	√
Coniferous	vegetation	Percentage of coniferous species in the pixel (%)	58.3 [0; 100]	√	√	√	√	

FIGURES

Fig. 1. Conceptual framework showing the interrelationships among the four main components that influence predictive errors in digital soil mapping when using statistical approaches. Getting the lowest prediction errors between observed and predicted soil properties on independent data is the main objective of digital soil mapping. Conceptually, the causes of prediction errors can be divided into four main components: (1) the quality and availability of the data (e.g., sample size, quality, spatial resolution and precision); (2) nature complexity or the level of heterogeneity in soil properties; (3) the choice of statistical framework (e.g., Bayesian vs frequentist), statistical method and algorithm, hereafter referred as 'statistical methods'; and (4) the choice of statistical model (e.g., spatial vs non-spatial, linear vs non-linear effects, simple vs interaction effect terms). Each of these components can act alone (bold arrows) or interact with other components (dashed arrows) to shape the accuracy of digital soil maps.

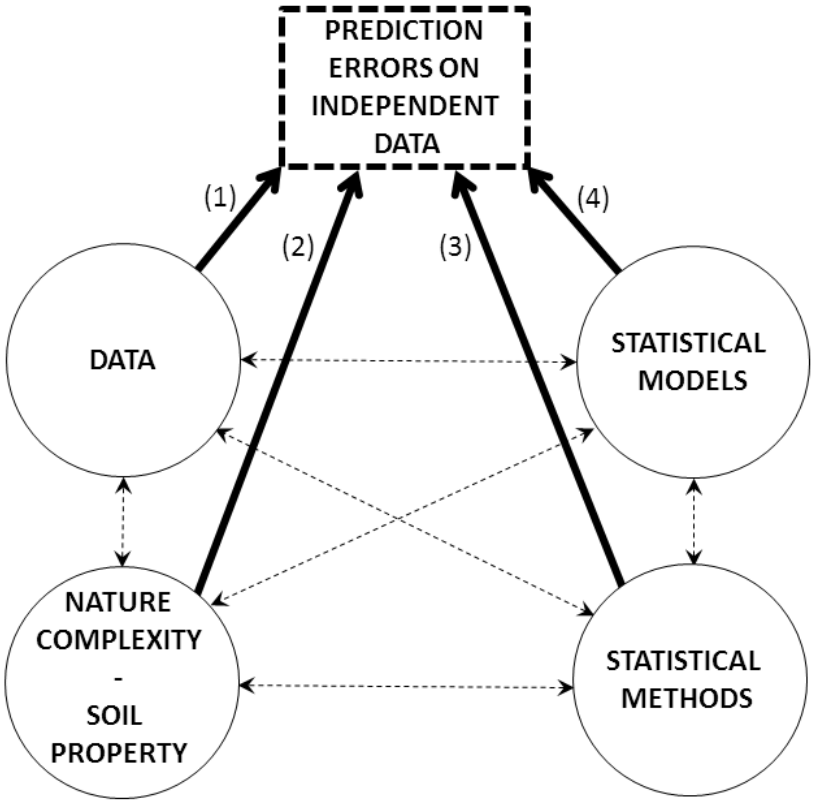

Fig. 2. Map showing the extent of the study area in Canada's managed forest (dark green) and the spatial distribution of soil profile data (black triangles). Soil profiles used as training data sets are shown as black triangles and soil profiles used for independent validation are shown as red triangles.

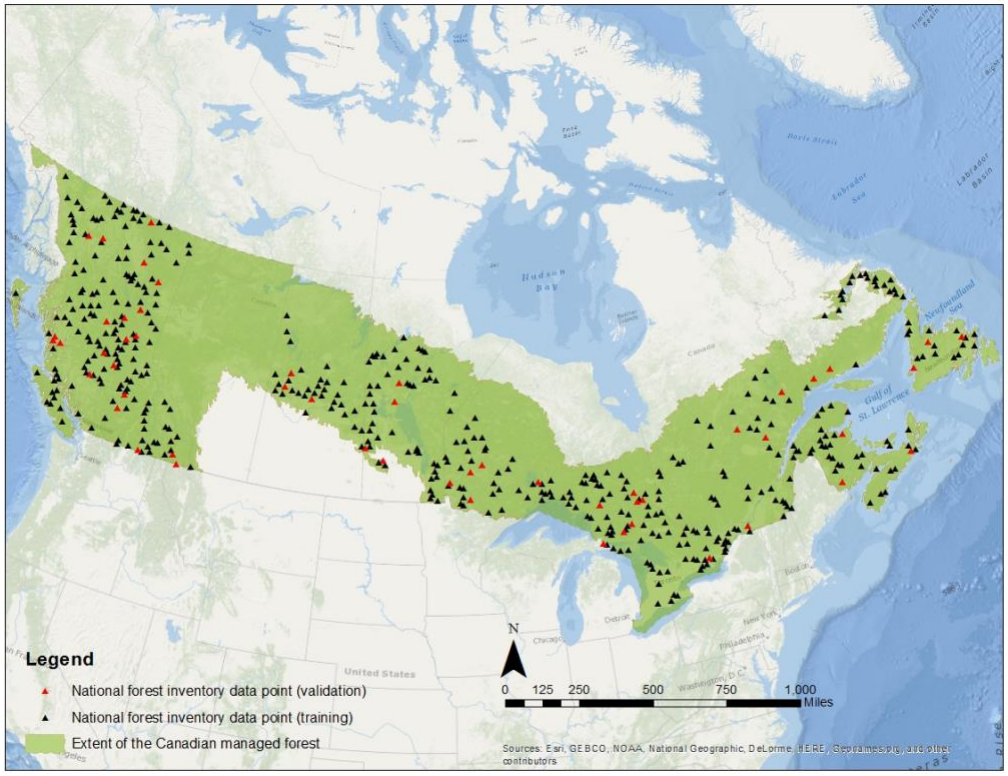

Fig. 3. Diagram showing the main steps of the modeling process used in this study. The 'model specification' step involves the selection of three different model types: 1) "*covariates only*": non-spatial models using only environmental covariates (topography, vegetation, and climate conditions) as predictors; 2) "*spatial only*": spatial models using only a function of the geographic coordinates of sample plots to predict soil properties. Note that each spatial function is different and specific to each statistical method (see Fig. 4); 3) "*covariates + spatial*": spatial models that combine both the effects of environmental covariates and a spatial function as above. Each of these three model types is fitted with every statistical method/framework (N = 8) at the 'selection of statistical method/framework' step. This process yielded a total of 24 combinations of model type-statistical method for each soil property. The predictive performances of each of the 28 combinations are compared using 10-fold cross-validation repeated 20 times. Predictions are then compared with values observed on independent data.

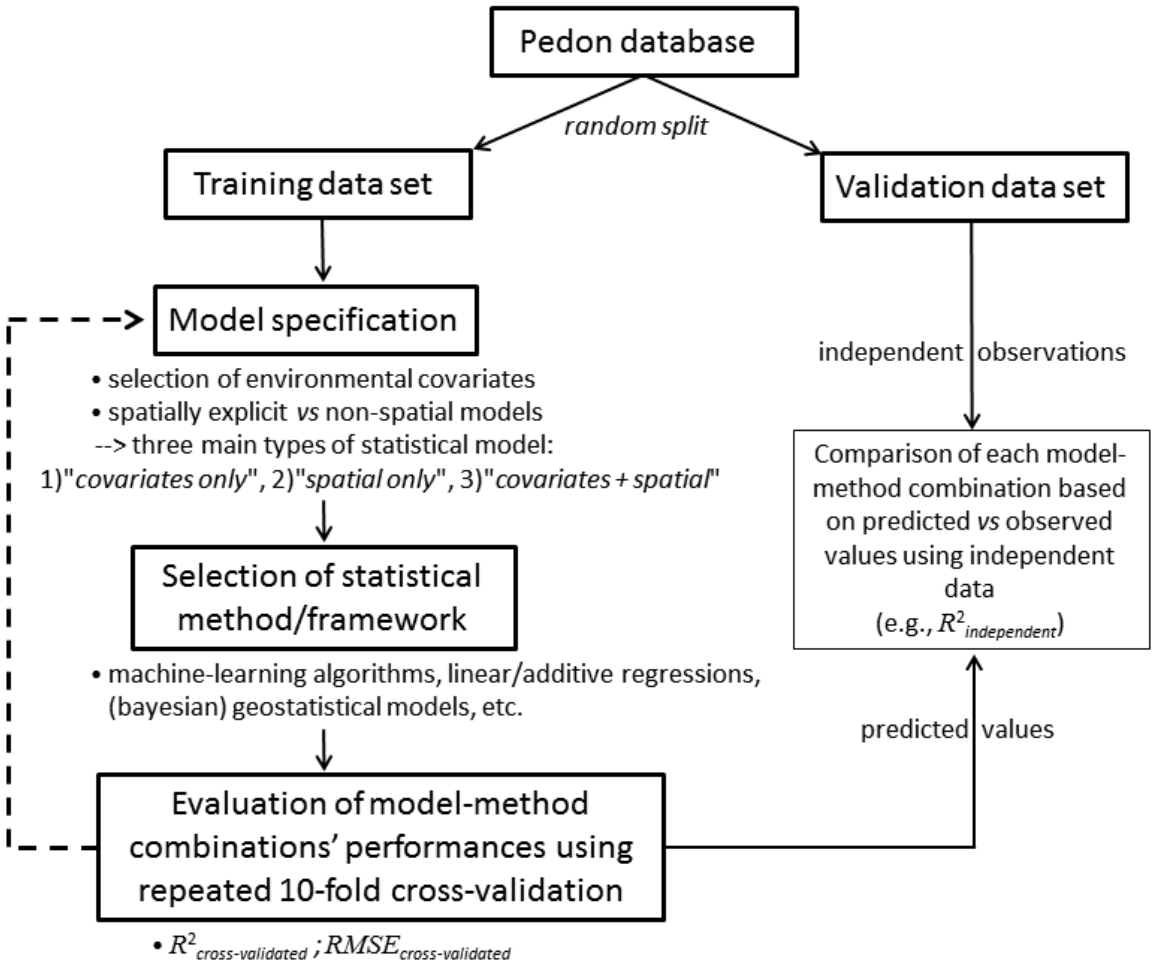

Fig. 4. Cross-validated R^2 and root mean square error (RMSE %) (median ± 95% quantile intervals) for five soil properties using 10-fold cross-validation repeated 20 times. Soil variables of the organic layer: carbon-nitrogen ratio (= C:N ratio organic) and thickness (cm) (= Thickness organic). Soil variables in the top 15 cm of the mineral horizon: bulk density (g/cm^3) (= Bulk density); carbon concentration (g/kg) (= C mineral), and the percentage of sand (= Sand mineral). Values are depicted as a function of the statistical method (y-axis) and type of model used (see colors in the legend). Note that for visual clarity, RMSE quantile values for Thickness organic and C mineral variables have been downscaled by a factor two and three, respectively (see right corner of each panel). Acronyms of statistical methods: INLA = integrated nested Laplace approximation; Kriging = kriging (ordinary or regression-kriging); GLM = generalized linear model; GAM = generalized additive model; Cubist = Cubist algorithm; KKNN = weighted *k*-nearest neighbors; BRT = boosted regression trees; RF = random forests; SPDE = stochastic partial differential equation approach.

(Figure 4. continued)

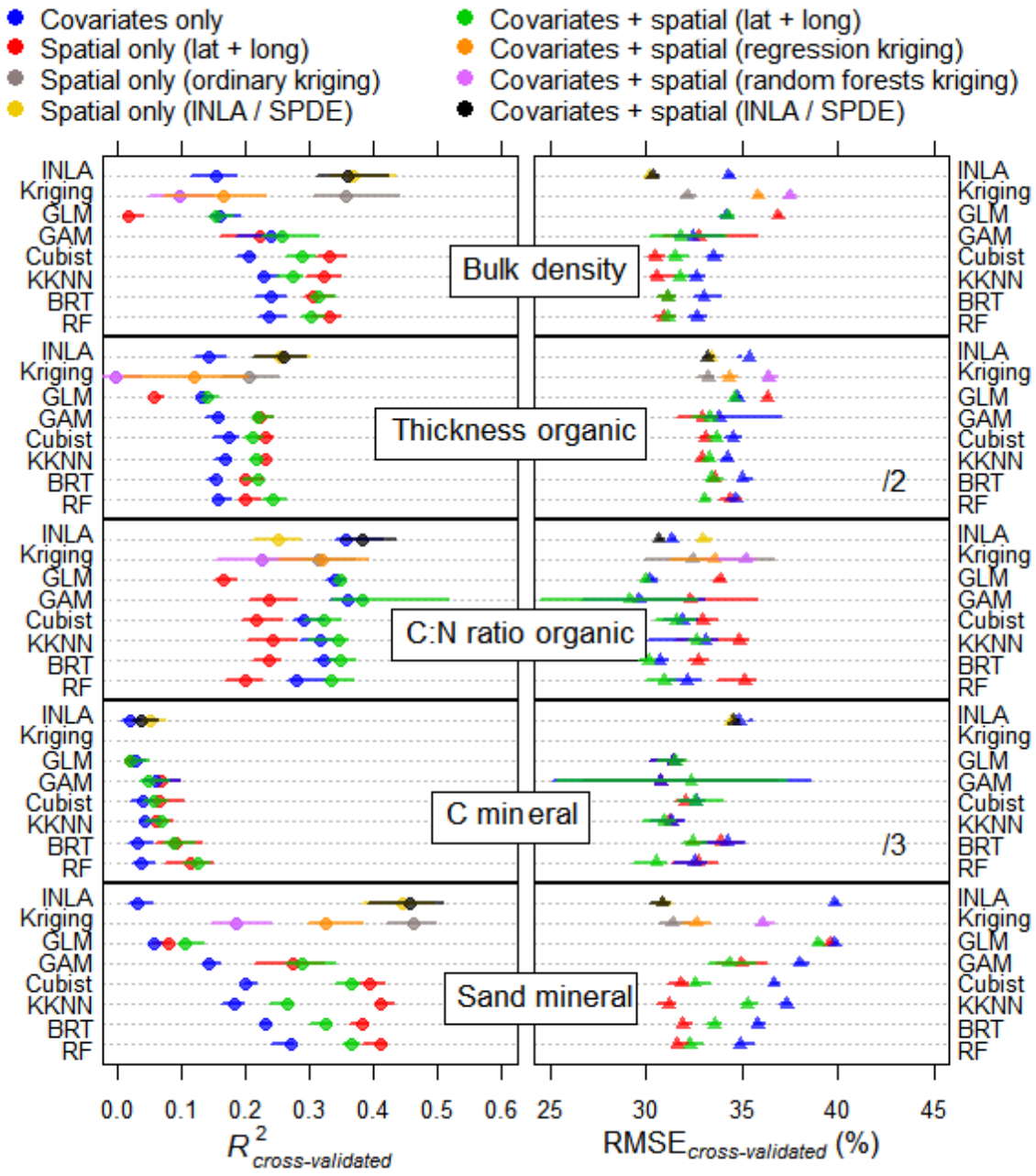

Fig. 5. Values of $R^2_{independent}$ (panel in the upper-left corner) and comparison between predicted and observed values of bulk density (all other panels) based on independent data as a function of model specification ("*covariates only*", "*spatial only*", and "*covariates + spatial*") and the statistical method used. Crosses and points' colors are identical to those in Fig. 3. The black line represents a 1:1 relationship.

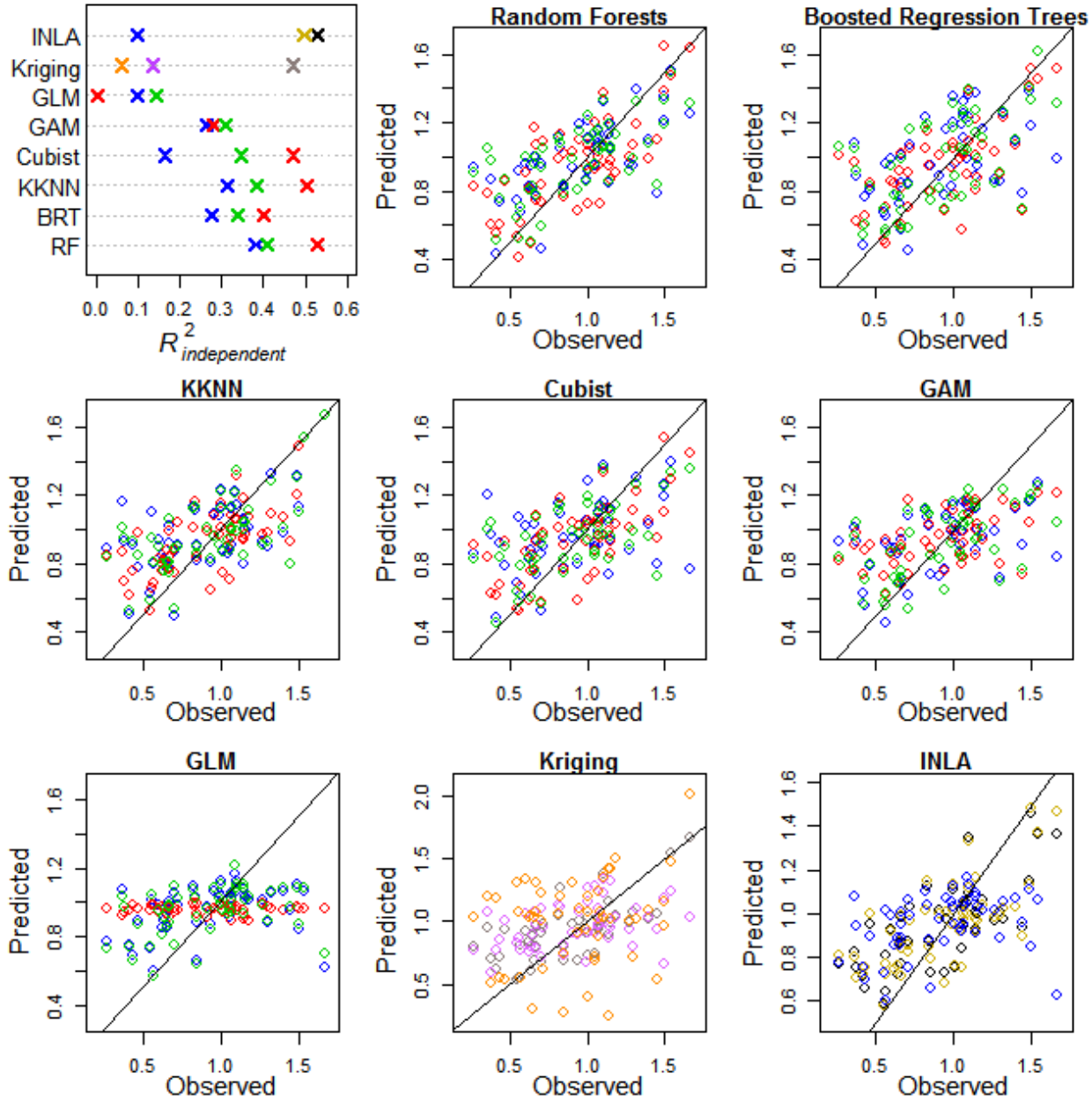

Fig. 6. Predicted soil properties (mean + standard deviation) obtained with the best model ("*covariates + spatial*") fitted with INLA (see Table 2 and Fig. 3). Left-hand side panel: predicted posterior mean for (A) sand content (%) in the top 15 cm of the mineral horizon; (B) C:N ratio in the organic layer; and (C) bulk density in the top 15 cm of the mineral horizon. Right-hand side panel: posterior standard deviation (= uncertainty maps) of the same variables as in the left-hand side panel (D, E, F).